Precision-Recall-Classification Evaluation Framework: Application to Depth Estimation on Single Images

Guillem Palou Visa and Philippe Salembier

Technical University of Catalonia, Barcelona, Spain

Abstract. Many computer vision applications involve algorithms that can be decomposed in two main steps. In a first step, events or objects are detected and, in a second step, detections are assigned to classes. Examples of such "detection plus classification" problems can be found in human pose classification, object recognition or action classification among others. In this paper, we focus on a special case: depth ordering on single images. In this problem, the detection step consists of the image segmentation, and the classification step assigns a depth gradient to each contour or a depth order to each region. We discuss the limitations of the classical Precision-Recall evaluation framework for these kind of problems and define an extended framework called "Precision-Recall-Classfication" (PRC). Then, we apply this framework to depth ordering problems and design two specific PRC measures to evaluate both the local and the global depth consistencies. We use these measures to evaluate precisely state of the art depth ordering systems for monocular images. We also propose an extension to the method of [2] applying an optimal graph cut on a hierarchical segmentation structure. The resulting system is proven to provide better results than state of the art algorithms.

Keywords: Precision-Recall, Detection, Classification, Depth ordering.

1 Introduction

While humans are very effective at estimating the scene structure from monocular images or sequences, computers are still very limited for this task. Many systems have been proposed but, performances of current unsupervised systems cannot compete with human perception. The work [4] stated that low level depth cues could be used to retrieve a global depth order. Although humans use these cues, their reasoning is also based on high order statistics and a priori knowledge on the type of scene. Nevertheless, low level cues do offer a good starting point to determine depth order. Works such as [3,17] attempt to estimate the depth order through the explicit detection of occlusion cues and rely on two perceptual ideas: 1) convex regions appear to be occluding and 2) in case of T-junctions, the region forming the largest angle is the occluding region. [11] proposes an extension of the gPb algorithm [12] to provide a figure/ground order based on the convexity of detected contours. More recently, the work in [2] retrieves the depth order by computing the *probability of ownership* of a pixel to different components. The algorithm supposes that the image is generated by a dead leaves model [8]. The novelty of this approach is that it does not have to explicitly deal with cue detection but occlusion arises naturally from the image model.

D. Fleet et al. (Eds.): ECCV 2014, Part I, LNCS 8689, pp. 648–662, 2014.

Table 1. Performance measures of [17,9,18] using figure/ground classification accuracy. All measures are extracted from the respective papers. Segmentation information is unavailable.

Method	BPT+TJC [17]	UCM+TJC [17]	[9]	[18]
Accuracy (%)	71.3	69.3	69.1	68.9

Other approaches rely on higher level features such as surface orientation or semantics. In this research line, the work of [20] oversegments the image and infers depth maps using a random field. In [6], the surface layout detector [5] and other features are used to detect the orientation and type of surfaces present in the image to condition a posterior inference on their spatial position. These approaches heavily rely on a training step and suffer when the type of scene has not been observed during the training phase.

The problem is also related to figure/ground (f/g) assignment on contours. In this field, [18,10,6,9] assign a depth gradient to each detected contour. The difference between f/g and depth ordering is that in the former, closed contours and regions are not necessary. Depth ordering, on the other hand, produces an image partition and a global depth interpretation. Conversion from depth ordering to f/g is possible by computing the depth gradient of the produced partition (but the converse is in general not possible).

Assessing the performance of f/g systems is traditionally done by measuring the f/g accuracy on detected contours. Table 1 reports the performance of several methods: [17], [9] and [18]. The usual way to measure the performances is to decouple segmentation from depth classification by providing two independent measures. Thus, the final f/g score is the classification accuracy of the boundary recall. The main problem with this approach is that it completely ignores the quality of the segmentation, leading to biased results if only confident contours are detected. However, the f/g performance is generally strongly related to the segmentation quality. As stated in [13], the f/g assignment on confident contours is easier than the assignment on ambiguous ones. Therefore, if a system only provides the most confident contours, the f/g score will be biased towards high values. [18] or [10] show results on both human marked and automatically detected contours and, precisely, much better f/g scores can be obtained with perfect segmentations. In other words, there exists a compromise between the segmentation quality and the f/g labeling problem which, to this day, has not been fully addressed.

In [13], a first step is proposed by evaluating the f/g score versus the boundary recall for video frames, showing that, effectively, there exists a compromise between these two values. However, this approach loses the precision information and thus does not provide a complete evaluation. For instance, Fig. 1 shows an image with its ground-truth depth order along with four possible outcomes of four different depth ordering systems. Which one is the best? The answer is not simple, as the user may sacrifice some segmentation quality so as to obtain correct depth relations or vice versa. Therefore, the question that naturally arises is to know whether it is possible to evaluate at the same time precision, recall and classification accuracy. We show that using a precision-recall-classification (PRC) framework, it is possible to provide both contour detection and depth gradient classification in a single plot and provide a complete view of the algorithm performances.

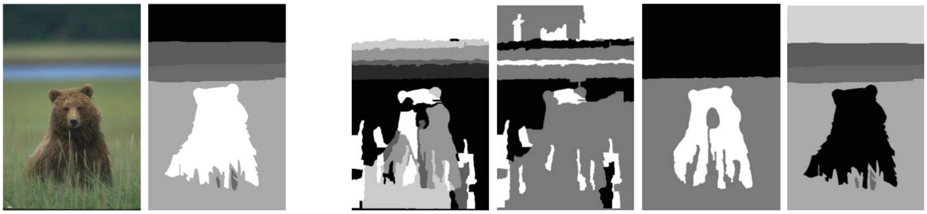


Fig. 1. From left to right: original image, ground-truth depth order and four depth order results. A part from the second result, deciding which is the best result is a difficult task.

"Detection plus classification" problems arise in many fields. For example, structured prediction with latent variables [7] classifies objects into classes without knowing their localization. In these problems, latent variables (which correspond to the detection step) are not explicitly modeled but they are key to the performance of the system. In [7], the latent variable is a bounding box indicating the object localization. The system output is the classification of the detected objects into specific classes (human, animal...). Therefore, as for depth ordering, the problem involves the same two steps: 1) detection of object localization and 2) object classification. Most of the time, the better the object localization is, the better will the classifier perform. As detection and classification performance are not independent, it is interesting to have an evaluation framework capable of capturing all the information.

To this end, a Precision-Recall-Classification (PRC) framework is proposed in this paper and two particular instantiations for depth ordering are discussed. The main idea is to combine the detection problem (segmentation quality) and the classification (f/g accuracy) into a single evaluation framework, showing that it is a more accurate and appropriate way of assessing performances than relying on two different measures on segmentation and classification. Besides this contribution, we also propose a new depth ordering system extending the work of [2] to integrate high quality regions. Furthermore, we also publish new annotations of the BSDS500 Dataset [1] involving depth order ground-truth (available at *http://imatge.upc.edu*).

The paper is organized as follows. Sec. 2 discusses the detection plus classification evaluation problem and defines the PRC framework. Sec. 3 proposes two measures to evaluate depth ordered partitions and, in Sec. 4, an extension of the method of [2] is discussed. The new depth ordering annotations for the BSDS500 dataset along with the experimental results are discussed in Sec. 5. Finally, conclusions are reported in Sec. 6.

2 The PRC Evaluation Framework

2.1 Detection Performance Measures

In detection problems, systems are designed to decide whether a given event or feature is present or absent in a given space. Given a ground-truth annotation, the ideal system behavior is to detect all possible entities without giving any false alarms. Quantifying a system performance is normally done by combining True/False Positives/Negatives to

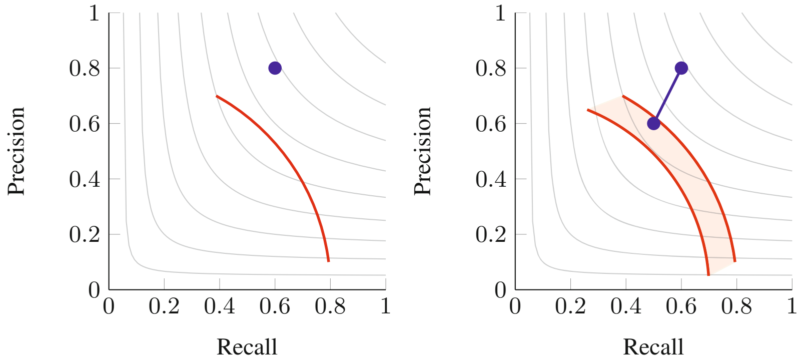


Fig. 2. Operating regions or points of two algorithms (red and blue) for the classic Precision-Recall (left) and the Precision-Recall-Classification (right) frameworks. Gray lines indicate points with the same F measure.

measure the *Precision* and *Recall*. Precision measures the rate of true positives among all detections, while Recall measures the percentage of detected ground truth annotations. They are defined by:

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN} \tag{1}$$

The ideal system corresponds to precision and recall equal to one. In practice, a compromise between these two quantities exists: a system with a high recall is likely to have false positives, and a system with high precision is likely to miss some true annotations. Often, the two quantities are summarized into a single number, F, defined as the harmonic mean of precision and recall: $F = \frac{2PR}{P+R}$.

Generally, the system performance is plotted on a precision-recall plane. As most systems depend on a given set of parameters θ, the precision and recall values also depend on θ: $P(\theta)$ and $R(\theta)$. This generates a set of points on the precision recall plane which are generally represented as a curve, see Fig. 2.

2.2 Combining Detection and Classification

In classification problems, results are often represented with confusion matrices, where the miss-classification rate is observed among different classes. If ground-truth results are available, the classifier performance can easily be computed. However, if classified objects should be first detected by an algorithm, it is likely that the classification score will depend on the operating point of the detection system. For instance, if only confident detections are considered (low recall, high precision), a high classification score is likely to be obtained. On the contrary, if many detections are retrieved, (low precision, high recall), the classification performance is likely to be worse. To integrate the detection and classification problems, we introduce two concepts:

- Inconsistent Detection ID: a correct detection that has been erroneously classified.
- Consistent Detection CD: a correct detection that has been properly classified.

Table 2. Confusion matrix of the proposed PRC framework. $\emptyset$ indicates no detection, while 1 indicates a detection. A and not A are the possible outcomes of the classifier. TN, MD and FD stand for true negatives and missed detections. The other concepts are defined in the text.

		Detection: $\emptyset$	Detection: 1	
			Class: A	Class: not A
Detection: $\emptyset$		TN	MD	MD
Detection: 1	Class: A	FD	CD	ID
	Class: not A	FD	ID	CD

All possible combinations of system output and ground-truth annotations are shown in Table 2. Similarly to pure detection scores, these measures are combined to provide precision-recall measures. CD and ID should be interpreted with care. Note that IDs, although not desirable, are in some way "better" than miss-detections MD or false detections FD since a correct detection is present and a post-processing step may correct the classification. Let us consider two extreme cases of evaluation:

Pure Detection System. In this scenario, we ignore the classification and consider an outcome to be correct if the detection is correct. In this approach CD and ID are equivalent and $TP = CD + ID$, $FP = FD$ and $FN = MD$.

Pure Classification System. This scenario considers that an outcome is correct if and only if detection and classification are correct. Hence, one should consider that TP are only correctly detected events with the same classification as the ground-truth. In this context, $TP = CD$ while ID should be interpreted in two ways:

- Detecting an incorrect class is equivalent to detect an event/object that does not exist. Therefore $FP = FD + ID$.
- Detecting an incorrect class leaves a ground-truth result without correct detection. Therefore $FN = MD + ID$.

To consider a scenario in-between these two extremes, a parameter $0 \leq \beta \leq 1$ is introduced to define the compromise between segmentation and classification qualities. In this way, it is possible to redefine:

$$TP(\beta) = CD + \beta ID \tag{2}$$

$$FP(\beta) = FD + (1 - \beta)ID \tag{3}$$

$$FN(\beta) = MD + (1 - \beta)ID \tag{4}$$

Precision (P) can be redefined using (1), (2) and (3):

$$P(\beta) = \frac{CD + \beta ID}{CD + \beta ID + FD + (1 - \beta)ID} = \frac{CD + \beta ID}{CD + ID + FD} = C_p + \beta I_p \tag{5}$$

With $C_p = \frac{CD}{CD+ID+FD}$ and $I_p = \frac{ID}{CD+ID+FP}$ which are the consistent and inconsistent precision respectively. Similarly, the recall (R) can be redefined as:

$$R(\beta) = \frac{CD + \beta ID}{CD + ID + MD} = C_r + \beta I_r \tag{6}$$

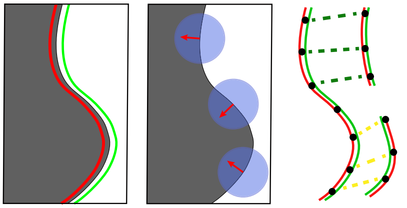


Fig. 3. Left: Depth partition with one contour. The green (red) overlay indicates the figure (ground) side. Center: Contour normals are estimated by averaging local orientations. Right: Bipartite matching of the ground-truth contour (right) and detected contours (left). Consistent (green) and inconsistent (yellow) matchings are shown.

C_r and I_r are the consistent and inconsistent recalls. Therefore, as shown in Fig. 2, each operating point establishes a line segment on the precision-recall (PR) plane depending on β. If the algorithm depends on a set of parameters θ, the evaluation produces a region in the PR plane. To differentiate these measures with respect to a classical detection approach, we will refer to them as Precision-Recall-Classification (PRC) framework.

The PRC plot of Fig. 2 gives insight about the system performance. Ideally, a system should reach $P(\beta) = R(\beta) = 1$ for all β values. Real systems however present a compromise between precision and recall. In the PRC framework, there is an additional compromise corresponding to the width of the operating region. A wide region indicates poor system performance in classification ($I_p, I_r \gg 0$), while a thin region ($I_p, I_r \approx 0$) indicates that the system is a good classifier. Moreover, as the operating point of the detection system detects only confident event/objects (low recall), the region width is expected to decrease, as classification is easier. Based on this framework, concrete PRC measures are proposed in the next section.

3 PRC Depth Measures

Equations (5), (6) defines the abstract PRC framework without specifying CD and ID for a specific problem. In this section, we show how the PRC framework suits a depth ordering evaluation task by proposing two measures.

3.1 Local Depth Consistency

We extend here the original bipartite matching for contour detection evaluation [14] to include the classification step: when a contour is correctly detected, it may be consistent (CD) with the ground-truth depth order (correct f/g assignment to both sides) or inconsistent (ID). Originally proposed in [18], the performance of a f/g classification algorithm is to simply measure $fg = \frac{CD}{CD+ID}$. The original matching scheme [14] is therefore modified to measure inconsistent matchings as follows (see Fig. 3):

1. From the depth partition, figure and ground sides are identified by examining the depth of each region.

2. The orientation of the depth gradient is estimated by averaging contour normals within a local window.
3. Bipartite matching of ground-truth and detected contours: CD and ID are marked with green and yellow lines respectively. A matching is inconsistent if the orientation of the depth gradient exceeds a specified threshold (15^o).

Once CD and ID are defined, Equations 5 and 6 define the so-called Local Depth Consistency (LDC) as it measures local depth relations on contours.

F/G over Random Index. As previously mentioned, precision and recall curves are sometimes summarized by a single number, F. Here we define a similar number for the PRC framework. According to equations (5) and (6), precision and recall are divided into their consistent and inconsistent subparts. Consider a contour detection system S and two classification systems on the detections of S: S_i and S_r. Suppose that for S_i, depth gradients are assigned using some sort of reasoning; while, for S_r, the depth gradient is randomly assigned. Therefore, the classification performance of S_i is expected to be better than the one of S_r at every detection operating point.

Assume the operating point of S_i has a total of D detections with a given set of CD_i and ID_i with $D = CD_i + ID_i$. S_r uses the same detection system S so, detections are the same. If S_r assigns randomly the depth gradient on a contour, and as the depth gradient can have two directions, the chance of assigning a correct depth gradient is 50%. Therefore, $CD_r = ID_r = \frac{D}{2} = \frac{CD_i+ID_i}{2}$. It is possible to show with (5) and (6) that the precision, recall and F measure (P_r, R_r, F_r) of a random classification system are related to their counterparts of an "intelligent" classification system (P_i, R_i, F_i) with the same detection score by:

$$P_r = (1+\beta)P_i/2 \qquad P_r = (1+\beta)P_i/2 \qquad F_r = (1+\beta)F_i/2 \tag{7}$$

When $\beta = 1$, no misclassification is considered wrong, so all measures are essentially the same. On the contrary, when $\beta = 0$, only correct contours count for precision and recall. So in a random system, there is a 50% chance of getting a consistent detection. We have found that this is a reasonable assumption for partitions with no severe oversegmentation, as matched detected contours follow ground-truth boundary orientation. However, when a highly oversegmented partition is evaluated, contour orientation cannot be easily estimated and thus the accuracy ratio $fg = \frac{CD}{CD+ID}$ can be lower than 50%. See the results section for more details.

Therefore, assuming this baseline, precision, recall and F-measure are reduced to half with respect to an intelligent classification system. It is expected that a real system will behave better than pure random guesses, so this can represent a lower bound of the system performance. If the system depends on a set of parameters, the F measure is a function of two variables $F(\beta, \theta)$. Define $\theta_{max} = \arg\max F(1, \theta)$ and use it to find $F_{max} = F(1, \theta_{max})$ and $F_{min} = F(0, \theta_{max})$. It is then possible to define an over-random-index (ORI) as:

$$ORI = \max(0, \frac{F_{min} - F_{max}/2}{F_{max}/2}) \tag{8}$$

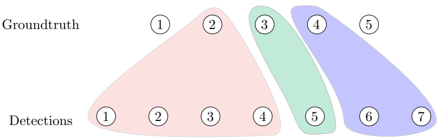


Fig. 4. Region matching example. Each detected region is matched to a ground-truth region. In case of subsegmentation, some ground-truth regions may not be matched. In case of oversegmentation, the same region may be matched multiple times.

When $ORI = 0$ the system behaves randomly, while when $ORI = 1$ the systems performs without misclassifications. Since ORI summarizes a whole region in the precision-recall plane into a single number, it only gives a rough indication of the system performance. The maximum operator is used to ensure positive ORI scores.

3.2 Global Depth Consistency

When estimating depth maps or figure/ground, it is important that the whole depth map is consistent with respect to a ground-truth. That is, the global depth image structure should be the same for the estimation and the ground-truth, even if the contours do not perfectly match. Therefore, a non local measure that quantifies the global depth consistency is desirable. To this end, similarly to the LDC, we have designed a region based precision-recall framework called Global Depth Consistency (GDC).

Assume the system output is a partition P_S formed by a set of regions $S = \{S_i\}$ and the ground-truth data is also a partition P_G with regions $G = \{G_i\}$. Unlike contours, regions by themselves do not incorporate the notion of relative order. However, if we consider pairs of regions, the notion of depth transition naturally arises. Since these pairs of regions do not necessarily need to be adjacent (unlike contours, which delimit two adjacent regions), evaluating all pairs of region order leads to a global depth interpretation of the estimated P_S with respect to P_G. Let Δ_i^S, Δ_i^G denote the depth of regions S_i, G_i. Prior to the evaluation, each region S_i is matched with a ground-truth region by finding its maximum Jaccard index (See Fig. 4):

$$m(S_i) = \widetilde{G}_i = \arg\max_{G_i} \frac{S_i \bigcap G_j}{S_i \bigcup G_j} \qquad \forall S_i, G_j \tag{9}$$

When matchings are available, FD (False Detection) is the number of detected pairs of regions matched to the same ground-truth region with different depth (See Fig. 5):

$$FD = \sum_{S_i, S_j \in S} \left(1 - \delta\left(\Delta_i^S, \Delta_j^S\right)\right) \delta\left(\widetilde{G}_i, \widetilde{G}_j\right) \tag{10}$$

where $\delta(a, b) = 1$ if $a = b$ and 0 otherwise. MD (Missed Detection) is the total number of missed transitions due to the region matching process (9). If the set of unmatched ground-truth regions is $\widetilde{G}$, the formal expression of MD is:

$$MD = |G|\left(|G| - 1\right)/2 - |\widetilde{G}|(|\widetilde{G}| - 1)/2 \tag{11}$$

	1	2	3	4	5
1	⬜	🟥	🟥	🟥	🟥
2	⬜	⬜	🟨🟩	🟨🟩	🟥
3	⬜	⬜	⬜	🟩	🟥
4	⬜	⬜	⬜	⬜	🟥
5	⬜	⬜	⬜	⬜	⬜

	1	2	3	4	5	6	7
1	⬜	🟦	🟦	🟦	🟩	🟨	🟩
2	⬜	⬜	🟦	🟦	🟩	🟩	🟨
3	⬜	⬜	⬜	🟦	🟨	🟨	🟨
4	⬜	⬜	⬜	⬜	🟨	🟨	🟨
5	⬜	⬜	⬜	⬜	⬜	🟩	🟩
6	⬜	⬜	⬜	⬜	⬜	⬜	🟦
7	⬜	⬜	⬜	⬜	⬜	⬜	⬜

Fig. 5. Tables showing all possible ground-truth pairs (left) and detected region pairs (right). Red squares count as MD, blue count as FD, green as CD and yellow as ID. See the text for an extended explanation.

where $|\cdot|$ denotes the set cardinality. CD and ID are found by examining each pair G_i, G_j and averaging the pairs of detected regions with the same and different depth order respectively. This is done to avoid counting the same ground-truth transition twice for two different pairs of S_i, S_j. Intuitively, CD and ID for a pair G_i, G_j measures how, in average, the detections are consistent with the ground-truth depth. Define α_{ij} and β_{ij} the number of consistent and inconsistent matches for a pair G_i, G_j respectively. $\gamma_{ij}^{G,S} = \text{sgn}(\Delta_i^{G,S} - \Delta_j^{G,S})$ is an indicator of the order of the regions i, j in the sets G, S. Then, α_{ij}, β_{ij} are given by:

$$\alpha_{ij} = \sum \delta\left(\gamma_{kl}^S, \gamma_{ij}^G\right), \qquad \beta_{ij} = \sum 1 - \delta\left(\gamma_{kl}^S, \gamma_{ij}^G\right) \tag{12}$$

Both summations are performed over the regions S_k, S_l fulfilling $m(S_k) = \widetilde{G}_i$ and $m(S_l) = \widetilde{G}_j$. The final consistent and inconsistent measures are given by:

$$CD = \sum_{G_i, G_j} \frac{\alpha_{ij}}{\alpha_{ij} + \beta ij} \quad ID = \sum_{G_i, G_j} \frac{\beta_{ij}}{\alpha_{ij} + \beta ij} \tag{13}$$

The GDC is more restrictive than LDC because it considers not only local relations but also non-adjacent depth transitions. Therefore it is expected that precision-recall values be lower than in the LDC measure as a global consistency is harder to attain than a local one. These two measures will be used in Sec. 5 to evaluate state of the art algorithms plus a new one that is presented in the following section and that is an extension of [2].

4 Depth through Probability of Ownership

The method of [2] estimates the relative depth by measuring the likelihood of a pixel to belong to different connected components, named Probability of Ownership (PO). The algorithm does not explicitly detect low level cues, but estimates the likelihood that a pixel $\boldsymbol{p}$ belongs to the connected components formed by a dead leaves model [8]. In this model, the image is assumed to be a superposition of objects of different sizes and depths. When objects are projected into the image plane, they are occluded by other objects of lower relative depth. If no occlusions were present, the projection of an object O_i would create a region X_i. When multiple objects occlude each other, the visible part of O_i becomes A_i ($A_i \subset X_i$) and the union of the visible parts A_i forms the image.

Although a pixel $\boldsymbol{p}$ belongs only to a visible component A_i, it may correspond to multiple (and occluded) X_i. The algorithm [2] exploits shape, distance and color features of A_i to determine a membership function, $Z(\boldsymbol{p})$, which estimates for each pixel $\boldsymbol{p}$ its probability to belong to several connected components of the image.

$$Z(\boldsymbol{p}) = \sum_{i=1}^{N} D(\boldsymbol{p}, A_i) \tag{14}$$

The density term $D(\boldsymbol{p}, A_i)$ estimates the likelihood that $\boldsymbol{p}$ belongs to X_i. It is defined using two principles: a pixel is more likely to belong to a set X_i if 1) the pixel is close to A_i and 2) the boundary is highly curved. The concrete expression of $D(\boldsymbol{p}, A_i)$ is rather complex and we refer the reader to [2] for more details. The function $Z(\boldsymbol{p})$ is an indicator of the number of X_i that a pixel belongs to. If $Z(\boldsymbol{p}) = 1$ the pixel belongs to a single component, while if $Z(\boldsymbol{p}) > 1$ the pixel belongs to more than one component. The only reason for a pixel to belong to more than one component is occlusion. Therefore, $Z(\boldsymbol{p})$ is a direct indicator of local depth without explicitly detecting occlusion cues. The higher $Z(\boldsymbol{p})$ is, the closer will be the pixel to the viewer. From now on, we will refer to $Z(\boldsymbol{p})$ values as the Probability of Ownership (PO).

4.1 Incorporating Regions

In practice, the sets A_i are formed by pixels with no semantic information. Therefore, the algorithm [2] is based on processing raw color pixel information. However, if higher level information, such as regions, needs to be extracted, the values $Z(\boldsymbol{p})$ can be used in conjunction with segmentation hierarchies to retrieve relevant objects. Here we propose to estimate the depth order map through an optimal graph cut applied on a hierarchical segmentation structure represented by a Binary Partition Tree (BPT) [19] which has been previously populated with PO values.

First, in order to construct the BPT, the ultrametric contour map (UCM) [1] of the image is computed. Then, the BPT is created by successively merging the pair of neighboring regions separated by the lowest salient contour as defined by the UCM. The leaves of the BPT correspond to the regions belonging to the finest UCM partition and the remaining BPT nodes represent regions obtained through the merging of pair of neighboring regions. The BPT root corresponds to the entire image support. Finally, the BPT edges describe the inclusion relationship between regions.

A partition can be naturally extracted from a BPT by selecting the regions represented by the tree leaves. If this is done on the original tree, the leaves correspond to the finest UCM partition and the process is trivial. However, if we prune the tree, that is if we cut branches at one location to reduce their length, a new tree, called a *pruned BPT* is created and the leaves of the pruned tree may define useful partitions. The pruning can be seen as a particular graph cut: Assume the tree root is connected to a *source* node and that all the tree leaves are connected to a *sink* node. A *pruning* is a graph cut that separates the tree into two connected components, one connected to the source and

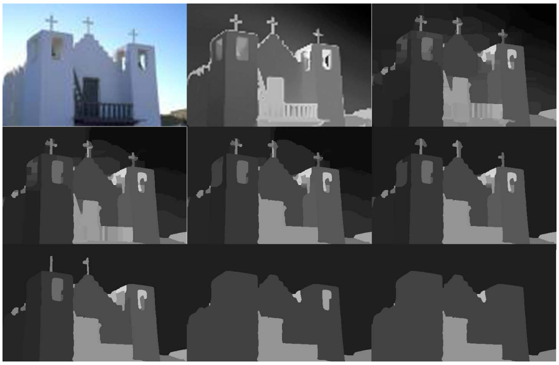


Fig. 6. Example of results using [2] and the proposed region-based approach. In raster order: original image, original depth map of [2] and depth maps found with increasing λ in Eq. (15). Note that the region-based system is able to resolve some error between the building and the sky.

the other to the sink, in such a way that any pair of siblings falls in the same connected component. The connected component that includes the root node is the pruned BPT and its leaves define a partition of the space.

To extract a depth order partition, the tree nodes, representing regions R_i, are populated by their mean PO value $\widehat{D}_i$ and their perimeter Γ_i. The pruning is defined by the graph cut minimizing a criterion inspired by the Mumford-Shah functional [15] over the hierarchy. If $\{R_i\}_{1\leq i\leq N}$ denotes the set of regions corresponding to the leaves of the pruned BPT, the criterion to optimize is given by:

$$\sum_{i=1}^{N}\left(\sum_{p\in R_i}\left|\widehat{D}_i - D(\boldsymbol{p})\right|^2 + \lambda|\Gamma|_i\right) \tag{15}$$

where $D(\boldsymbol{p})$ is the original estimated depth (PO) value for pixel $\boldsymbol{p}$ and λ a parameter controlling the partition granularity. The criterion defined by Eq. (15) can be very efficiently minimized by dynamic programming [19,21,16]. Small λ values create fine partitions with many regions, while for larger λ, coarser partitions are found. The final depth partition is formed by regions R_i with $\widehat{D}_i$ as depth value. Fig. 6 shows several partitions obtained with this optimum pruning.

5 Experiments and Results

In this section, we compare the classical depth ordering evaluation with the LDC and GDC measures. For the selection of the state of the art systems, the code of [18] and [9] are not public, so we could not run the algorithms for evaluation. Their performances as given in their respective papers are reported in Table 1. The LDC and GDC measures are reported for the BPT+TJC and UCM+TJC approaches of [17], the angular embedding (AE) of [11], the occlusion boundary detection (OB) of [6], the learning depth based (LD) approach of [20], the PO approach of [2] and the proposed algorithm using the region PO-based graph cut on BPT.

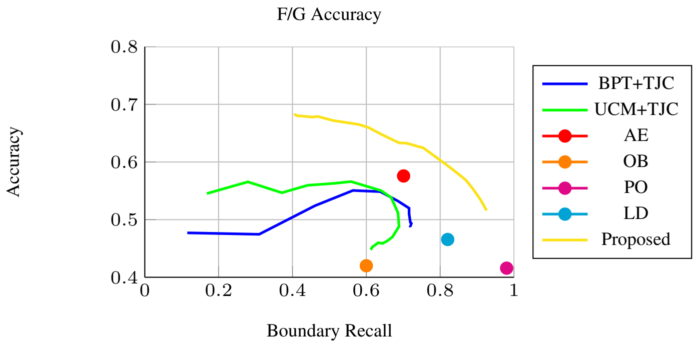


Fig. 7. Accuracy of the f/g classification as a function of the boundary recall

5.1 Dataset with Depth Annotation

We have found few public datasets incorporating relative depth ordering between objects present in images. One of the most popular datasets in image segmentation, the BSDS500 [1] incorporates figure/ground annotations for a subset of the images. Although it is the classical evaluation choice for figure/ground system, it does not involve closed contours and no global consistency is found in several cases. To solve this issue, we have chosen several segmentations for each of the 500 images and have annotated their regions with relative depth, creating a consistent depth map. Examples can be seen in the second column of Fig. 9. These annotations are publicly available at *http://imatge.upc.edu*.

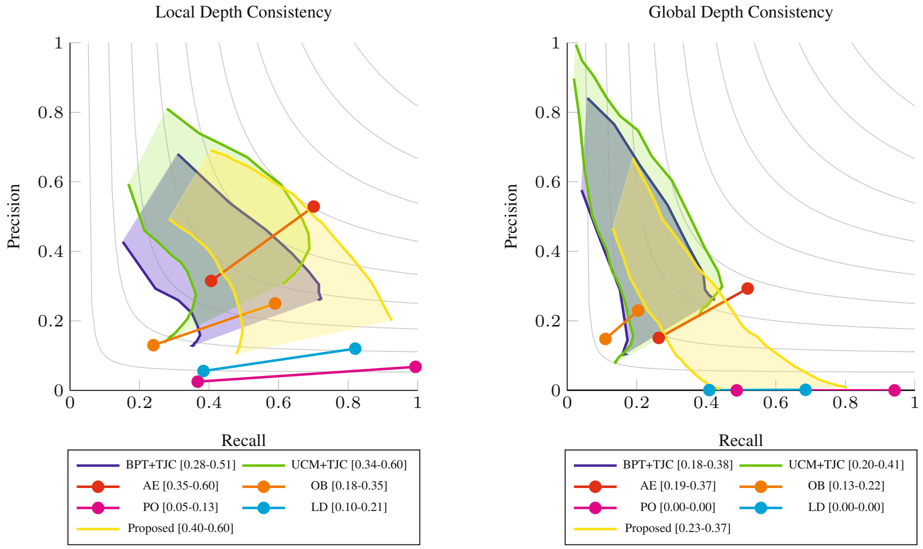
5.2 LDC and GDC Results

Fig. 7 shows the figure/ground classification accuracy as a function of the boundary recall[1]. Although this plot is an interesting step for evaluating depth ordering systems, it is not found in the literature other than in [13] for f/g estimation in video frames.

It follows from Fig. 7 that the proposed system seems the one providing more consistent depth estimates. Nevertheless, the f/g accuracy does not provide the segmentation precision, a crucial segmentation measure. Moreover, if precision is not known, depending on the degree of recall that one may require, the proposed system can present lower accuracy than BPT+TJC, UCM+TJC and AE. Therefore, additional measures should be invloved to define the quality of each system.

To this purpose, these systems are evaluated with the LDC and GDC measures. Results are presented in Fig. 8. The first point to notice is that LDC gives higher scores than GDC: a global depth consistency is much harder to achieve than estimating local depth gradient on contour points. Second, the LDC measure does suffer for output

[1] The f/g matching process can lead to different results compared to the original papers, due to differences in contour distance and angle tolerance on the depth gradient. Following a strategy similar to [18], we only considered contours at a 3 pixels distance for matching. A depth gradient is considered correct if its orientation lies within 15° of the ground-truth orientation.



	BPT+TJC	UCM+TJC	AE	OB	PO	LD	Proposed
ORI	0.12	0.13	0.18	0.03	0.00	0.00	**0.32**

Fig. 8. LDC and GDC measures for state of the art system. F_{max} and F_{min} measures for each method are shown in brackets in both cases. The ORI is also shown in the table at the bottom.

with high recall values and low precision. The bipartite assignment for oversegmented solutions (such as the PO and LD systems) has problems to assign the depth gradient when many contours are potential matches. Since the assignment does not favor any particular orientation, the ORI index approaches 0 for high recall methods.

LDC: Overall, the system proposed in this paper is the best one in terms of segmentation quality and of depth ordering, as its ORI is more than 0.1 over the second best technique. Still, there is much room for improvement both in segmentation quality and depth ordering, since the maximum F measure is around 0.6, where the ground-truth assignments achieve an F measure of 0.78. The ORI index indicates that monocular depth ordering still can be improved, as the theoretical maximum ORI is 1 and the best kown technique has a much lower score (around 0.3).

GDC: Referring to a global depth interpretation, the precision values of the systems have much lower scores compared to LDC. Basically because small regions are missed unless very oversegmented solutions are considered. An important point to notice is that PO has the largest recall for classification, indicating that it is the technique detecting most depth transitions. Again, the small width of the proposed algorithm region indicates that it behaves quite well for global depth interpretation. However, techniques as UCM+TJC or BPT+TJC seem to provide higher quality regions. Overall, we can conclude that proposed system should be the choice if monocular depth ordering is needed due to its better scores in LDC, ORI and its competitive results in GDC.

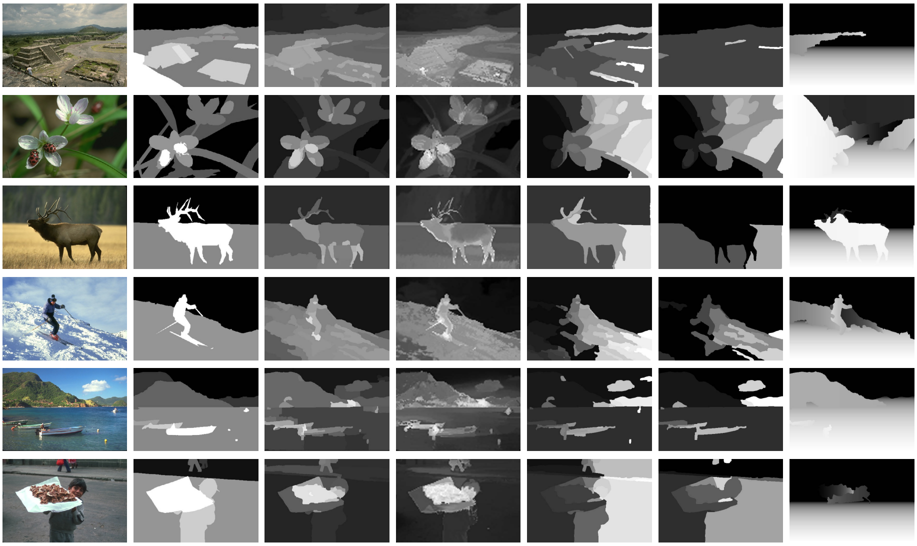


Fig. 9. From left to right. Original image, ground-truth relative depth order and results from the proposed system, PO, AE, the UCM+TJC and OB. The proposed methods is able to reduce noisy estimation of PO. More results can be found in the supplemental material.

6 Conclusions

We have proposed a new framework to evaluate problems encompassing detection and classification where the complete system performance can be seen in a single plot. Two particular applications of the PRC framework are shown for depth ordering, where local and global depth consistency measures are presented. Depth ordering annotations of the BSDS500 Dataset were also created for this particular problem and made public. State of the art methods and a new proposed algorithm are evaluated using both PRC measures. Results show that the proposed algorithmm is the one having the best results, although human depth perception is still unreachable for fully unsupervised systems.

References

1. Arbeláez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE TPAMI 33(5), 898–916 (2011)
2. Calderero, F., Caselles, V.: Recovering relative depth from low-level features without explicit T-junction detection and interpretation. IEEE IJCV (2013) (in Press)
3. Dimiccoli, M.: Monocular Depth Estimation for Image Segmentation and Filtering. Ph.D. thesis, Universitat Politecnica de Catalunya (2009)
4. Fowlkes, C.C., Martin, D.R., Malik, J.: Local figure-ground cues are valid for natural images. Journal of Vision 7(8), 2 (2007)
5. Hoiem, D., Efros, A.A., Hebert, M.: Recovering Surface Layout from an Image. IEEE IJCV 75(1), 151–172 (2007)

6. Hoiem, D., Efros, A.A., Hebert, M.: Recovering Occlusion Boundaries from an Image. IEEE IJCV 91(3), 328–346 (2011)
7. Kumar, M.P., Packer, B., Koller, D.: Self-paced learning for latent variable models. In: Advances in Neural Information Processing Systems, pp. 1189–1197 (2010)
8. Lee, A.B., Mumford, D., Huang, J.: Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model. IEEE IJCV 41(1-2), 35–59 (2001)
9. Leichter, I., Lindenbaum, M.: Boundary ownership by lifting to 2.1D. In: IEEE ICCV, pp. 9–16 (2009)
10. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: IEEE CVPR, pp. 1253–1260 (2010)
11. Maire, M.: Simultaneous Segmentation and Figure/Ground Organization Using Angular Embedding. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 450–464. Springer, Heidelberg (2010)
12. Maire, M., Arbelaez, P., Fowlkes, C., Malik, J.: Using contours to detect and localize junctions in natural images. In: IEEE CVPR, pp. 1–8 (2008)
13. Maire, M.R.: Contour Detection and Image Segmentation. Ph.D. thesis, University of California, Berkeley (2009)
14. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE TPAMI 26(5), 530–549 (2004)
15. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Comm. on Pure and Applied Mathematics 42(5), 577–685 (1989)
16. Palou, G., Salembier, P.: Depth ordering on image sequences using motion occlusions. In: IEEE ICIP, Orlando, FL, USA (2012)
17. Palou, G., Salembier, P.: Monocular Depth Ordering Using T-junctions and Convexity Occlusion Cues. IEEE Trans. on Image Proc. (2013)
18. Ren, X., Fowlkes, C.C., Malik, J.: Figure/Ground Assignment in Natural Images. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 614–627. Springer, Heidelberg (2006)
19. Salembier, P., Garrido, L.: Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. IEEE Trans. on Image Processing 9(4), 561–576 (2000)
20. Saxena, A., Ng, A., Chung, S.: Learning Depth from Single Monocular Images. In: IEEE NIPS, vol. 18 (2005)
21. Serra, J., Kiran, B.R., Cousty, J.: Hierarchies and Climbing Energies. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 821–828. Springer, Heidelberg (2012)